IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 26, 2019, accepted July 8, 2019, date of publication July 11, 2019, date of current version July 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2928125

WGAN-Based Robust Occluded Facial Expression Recognition

YANG LU, SHIGANG WANG, WENTING ZHAO, AND YAN ZHAO, (Member, IEEE)
College of Communication Engineering, Jilin University, Changchun 130012, China

Corresponding author: Shigang Wang (wangshigang@vip.sina.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0404800, in part by the National Natural Science Foundation of China under Grant 61631009, and in part by the Fundamental Research Funds for the Central Universities under Grant 2017TD-19.

ABSTRACT Research on facial expression recognition (FER) technology can promote the development of theoretical and practical applications for our daily life. Currently, most of the related works on this technology are focused on un-occluded FER. However, in real life, facial expression images often have partial occlusion; therefore, the accurate recognition of occluded facial expression images is a topic that should be explored. In this paper, we proposed a novel Wasserstein generative adversarial network-based method to perform occluded FER. After complementing the face occlusion image with complex facial expression information, the recognition is achieved by learning the facial expression features of the images. This method consists of a generator G and two discriminators D_1 and D_2. The generator naturally complements occlusion in the expression image under the triple constraints of weighted reconstruction loss l_{wr}, triplet loss l_t, and adversarial loss l_a. We optimize the discriminator D_1 to distinguish between real and fake by constructing an adversarial loss l_a between the generated complementing images, original un-occluded images, and small-scale-occluded images based on the Wasserstein distance. Finally, the FER is completed by introducing classification loss l_c into D_2. To verify the effectiveness of the proposed method, an experimental analysis was performed on the AffectNet and RAF-DB datasets. The visual occlusion complementing results, comparison of recognition rates of facial expression images with and without de-occlusion processing, and T-distributed stochastic neighbor embedding visual analysis of facial expression features all prove the effectiveness of the proposed method. The experimental results show that the proposed method is better than the existing state-of-the-art methods.

INDEX TERMS Facial expression recognition, partial occlusion, image complementation, Wasserstein generative adversarial network.

I. INTRODUCTION

Facial expression is a form of non-verbal communication, which is the main means of expressing social information between human beings. The facial expression recognition (FER) is a technology for using computer to recognize changes in facial features of a human face to classify different facial expressions. FER is not affected by race, skin, age, and gender [1]–[3]. Moreover, the expression recognition technology is a cross product of many disciplines, such as biology, psychology, and computer science. In recent years, this technology has gained significant attention of the researchers because of its high usage value and research significance in realizing the intelligence of human life. Furthermore, FER has become an important way of human communication because of the significance of facial expressions in human emotional expressions and the rich details it contains [4]. In addition, FER is widely used in areas, such as fatigue driving detection [5], [6], robotics [7]–[9], intelligent emotional computing [10], [11], and electronic classrooms [12]. The artificial intelligence and pattern recognition technology that has been highly valued and developed in recent years, because of which FER has been valued by more and more researchers as the intersection of the two.

Nowadays, human life is being incorporated with intelligence, and the key problem to be solved in intelligent life is to explore the true inner world and emotional expressions of human beings. Facial feature extraction in

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Asikuzzaman.

FER integrates digital image processing [13], [14], biology [15], [16], and statistical theory [17], [18]. Firstly the features of computer-processed facial expression images are extracted and represented by statistical models to complete modeling of expression feature. Then, the established expression feature model is compared with the testing expression feature, and the similarity between the two is measured to determine the category the expression belongs to, thereby completing the recognition of the facial expression by the machine.

FER is a challenging subject because it is an interdisciplinary technology, and the research and development of FER can promote both the theoretical significance and life applications. Currently, most of the related works of this technology is to identify un-occluded facial expression images, and the excellent research results are endless. However, in real life, the facial expression images captured by the image acquisition device often have partial occlusion [19], which commonly include obstructions from hands, glasses, masks, and so on [20]. These occlusions can interfere with the extraction of expression features and affect the accuracy of expression recognition. A system that can accurately recognize expressions under occluded conditions is for the need of the hour. With problems, such as illumination and noise, being solved one after another, researchers believe that a truly robust recognition method should be able to solve the problem of expression recognition under occlusion.

For the above analysis, in this study, we build an occlusion FER method based on the improved generative adversarial network, which can realize occluded face images complementation and FER. The method proposed in this study is based on the Wasserstein Generative Adversarial Network (WGAN) model, which consists of one generator and two discriminators. Under the triple constraints of the weighted reconstruction loss function l_{wr}, the triplet loss function l_t, and the adversarial loss function l_a, the generator makes full use of the pixel information of the un-occluded area in the occlusion image, and the structural information between the occlusion, generated, and original un-occluded images to fill the input occlusion expression image. The two discriminators corresponding to the adversarial loss function l_a and the classification loss l_c function are used to determine if an image is real/fake for expression classification respectively.

The remainder of this paper is organized as follows. In Section 2, we describe the structure and function of the occluded FER system and introduce the related research works of this technology. In Section 3, we introduce the occluded FER method proposed in this study and perform theoretical analysis from two aspects, which includes WGAN-based occluded FER system structure and the proposed four loss functions. In Section 4, we describe and analyze the experimental setups, experimental results, and comparative experiments. Finally, in Section 5, we conclude the paper and mention the scope for future improvements.

II. RELATED WORK

Generally, a complement framework for FER [21] consists of three parts, face detection and location [22]–[25], facial expression feature extraction [26]–[28], and expression recognition [29]–[31]. However, it is unrealistic to directly recognize the occluded expression images, because the occlusion area may be the eyes, mouth, or nose. These areas contain abundant expression features, and if the de-occlusion process is not performed, the key expression feature information will be lost, and the subsequent recognizing will not be performed smoothly. Therefore, occluded FER should first complement the occluded area [32], and then expression recognition is performed according to the complemented expression images.

A. OCCLUDED IMAGES COMPLEMENTATION

Partial occlusion facial expression recognition has attracted wide attention of researchers in recent years [70]. Currently, the methods used for complementing occluded image can be divided into two categories: discarding [33] and filling [34] methods. This study does not consider the discarding method, because it may lead to a lack of key expression features. In contrast, the filling method complements the occluded area of the image by measuring the pixel information of the occluded and un-occluded areas in the image, along with the hidden structure information of the occluded areas and the whole image. The methods in Refs. [35] and [36] were some traditional filling methods.

From image generation, face image complementation can be understood as a learning problem of probability distribution. There is contextual semantic association between image pixels. That is, the value of each pixel can be considered as the sample in the image probability space, and the filled image should be consistent with the original un-occluded image between facial expression and context. The most representative one is convolution neural network (CNN). Pathak *et al.* [37] proposed a context encoder-based image complementation method in 2016. They built a neural network of encoder-decoder structure, and inferred the information of the occluded area from the un-occluded area of the occluded images, to improve the quality of the filled image. That method added the discriminant loss of the authenticity of the generated image on the basis of the pixel reconstruction loss. Reference [37] demonstrates its superiority from two aspects. One is that the contextual encoder proposed by the author has a good ability to fill in the missing areas. On the other hand, the authors use the context encoder as a pre-training step for image classification, target detection and semantic segmentation. The results show that the learned characteristics can be transferred to other tasks. Experimental results show that the method proposed in [37] is superior to other unsupervised or self-supervised methods. Furthermore, Satoshi *et al.* [38] added a discriminator to the filling area based on Pathak's results. Different from [37], [38] used global and local context discriminators to train image completion networks. The whole image was judged by the

global discriminator to evaluate whether it was consistent. Meanwhile, the local discriminator is used to discriminate the completed areas to ensure the local consistency of the generated patches. The method proposed in [38] can not only maintain the local and global consistency of the image, but also optimize the details of the filling area. In addition, Yu *et al.* [39] used the convolution method to calculate patch similarity to complete the image inpainting task. The innovation of [39] was that the author proposes a new contextual attention layer to extract the features of the approximate area to be repaired from the remote area, which not only makes programming easier, but also reduces the calculation time of similarity.

In addition to CNN, the generative adversarial networks (GAN) is also applied in the field of image generation. Xu *et al.* [71] used the optimized Deep Convolutional Generative Adversarial Networks (DCGAN) to iteratively remove the occlusion in face images. This method can overcome the instability in the traditional GAN training process. Similar to [71], Yeh *et al.* [72] also discussed image inpainting based on DCGAN. The authors used context and prior loss to find the closest image coding in the potential image manifold. This encoding was then passed through the generated model to infer the missing content of the image. It also brought new ideas to the image inpainting at the sametime.

B. FACIAL FEATURE EXTRACTION AND RECOGNITION

The accuracy of the FER system is mainly affected by the feature extraction and classification methods. Most of the current research works are based on these two aspects.

Currently, expression feature extraction is the most important part in the FER system. There is no accurate evaluation standard for the quality of expression feature extraction methods, because the accuracy of features depends on specific problems and applied scenes. For facial expression images, the feature extraction methods based on geometric [40] and apparent features [41], [44]–[47] are both conventional. Expression classification and recognition is the last step of the FER system; however, it is a key step. The goal of expression classification is to judge the similarity between the features of test images and a certain type of expression features in the training set, and select the type with the largest similarity as the output result. The Support Vector Machine (SVM) [41] and sparse representation-based classification [48] are two popular traditional machine learning methods.

The above traditional feature extraction and classification methods are very effective for small datasets. With the advent of the big data era , deep learning methods, such as artificial neural networks, have emerged. For example, Kim *et al.* [49] proposed a CNN method for FER to avoid the complex feature extraction process in traditional FER. They extracted the hidden features of the expression images by training the convolution kernels, and used maximum pooling to reduce the dimensions of the extracted features. Finally, the SoftMax classifier was used to classify the facial expressions of the test samples. The experimental results showed that CNN had better FER performance and generalization ability.

C. GENERATIVE ADERSARIAL NETWORKS

GAN [50] is a product of a combination of game theory and deep learning. It is an unsupervised probability distribution learning method, which can learn the distribution of real data and generate new data sets with high similarity. GAN is composed of generator and discriminator. The generator learns the distribution of real sample data and to generate the most realistic fake-data. The discriminator is essentially a double classifier, which needs to identify if the input data is a real sample or a fake data generated by the generator. Inspired by the zero-sum game, a method to make generators and discriminators compete with each other, iteratively optimize in the confrontation, and continuously improve their generating and discriminating abilities was proposed in [50]. This was done so that generators can estimate the distribution of real data, and the generated samples are more real. GAN can be widely used in image generation, image complementation, semantic segmentation, image super-resolution, and image de-noising. Zhu *et al.* [67] captured the long-distance information in the output semantic label map based on GAN, which can enhance the spatial continuity in the output label map. and has effectiveness compared with the existing methods. Compared with existing methods, it can provide auxiliary higher-order potential loss for segmentation model, so that segmentation model has the ability to correct higher-order inconsistencies. Radford *et al.* [51] combined GAN and CNN together to obtain deep convolution generative adversarial networks (DCGANs). They used DCGAN for image classification, among which the generator and discriminator are both CNNs, because of its strong feature extraction and classification ability. Thus, the image classification results are better. In order to solve the problem of incomplete loss function and incompetent structure in existing networks Zhu *et al.* [68] proposed a generative adversarial image super-resolution method through deep dense skip connections (GSR-DDNet). This method model the mapping across the low-quality and high-quality images in an adversarial way. Therefore, the problem of super-resolution image blurring and over-smoothing can be solved by this method.

GAN can be used for classification in addition to generating visually realistic images. The success of deep neural networks in the fields of image recognition and object classification largely depend on a large number of manually labeled training datasets. However, in many applications, such labeled data volumes often fail to meet the requirements of deep model training and adding unlabeled sample data into training can perform semi-supervised classification. Zhang *et al.* [69] proposed a GANS-based active semi-supervised learning method. This method can obtain a comprehensive perception of the entire data distribution by learners adversarial or cooperative manner. So, it can eliminate the impact of the inadequacy of labeled instances and the unbalance within various classes on the learning

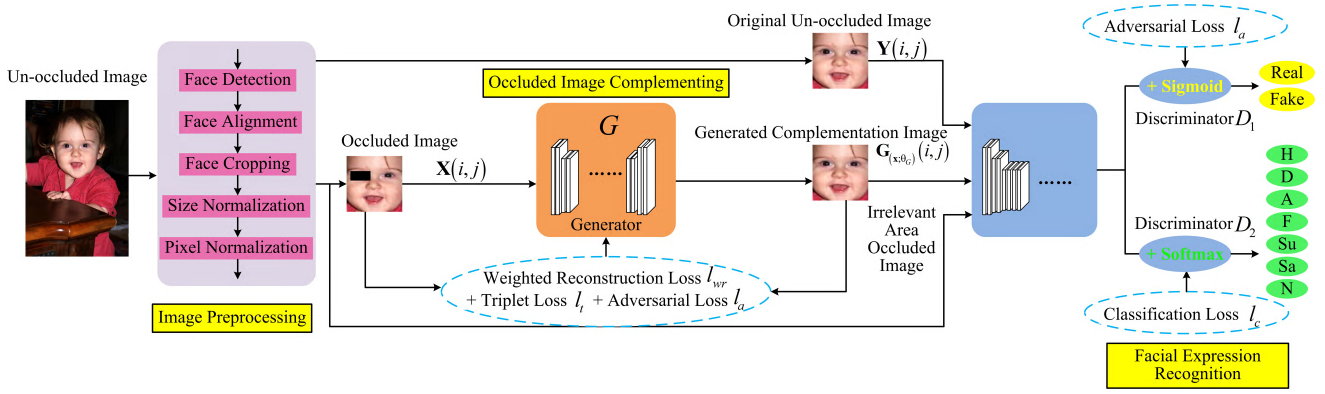


FIGURE 1. Framework of the robust occluded facial expression recognition based on the proposed method.

performance of the model. Zhu*et al.* [52] proposed to change the GAN discriminator from the original double classifier to a multi-classifier. At this time, the generator output samples can be used to train the classifier as the N + 1 class of the N classification problem. This model not only utilizes labeled training samples, but also learns characteristics from unlabeled generated samples.

It can be seen that GAN has been recognized by many researchers for its strong learning ability. In addition, some models are extended based on GAN, such as AE-GANs (Auto-Encoder Generative Adversarial Networks) [73], WGAN [56], etc. AE-GANs is a model which combines Auto-Encoder (AE) and GANs, so the advantages and disadvantages of AE and GANs can complement each other in order to improve the quality and stability of the generated image.The network structure of this paper is consistent with the idea of AE-GANs, because the input of this paper is an image, and the output is a generated one. However, we also use the network to classify seven kinds of expressions. At the same time, we introduce Wasserstein distance in WGAN network to measure the real/fake of the generated images.

III. PROPOSED METHOD

In view of GAN's advantages and characteristics, this paper based on it to recognize the occluded facial expression images. The method framework of this study is shown in Fig. 1, where G and D_o, $0 = 1, 2$, represent the generator and discriminators, respectively, $\mathbf{Y}(i,j)$ represents the preprocessed original un-occluded images, $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$ represents the generated complemented, $\mathbf{X}(i,j)$ represents the occluded images. The details of each part are described below.

A. OCCLUDED FER NETWORKS

Firstly, input the preprocessed occluded expression images $\mathbf{X}(i,j)$ into generator G, different from the original GAN, this study directly takes the occluded images instead of random noise as the input of the generator. G is a convolutional automatic encoder, which is composed of a pair of structurally symmetric encoder and decoder. The encoder carries out multiple convolution processing on the input image. Then the pixel information and the correlation information between the occluded area and the un-occluded area of the image are put into the fully connected layer. Afterwards, the decoder decodes these information to complement the occluded area of the input images. In this study, the proposed weighted reconstruction loss function and triplet loss function are used to constrain the generator together to achieve the purpose of optimizing image quality. The weighted reconstruction loss function l_{wr} is to make the generated complemented $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$ more similar to the original un-occluded image $\mathbf{Y}(i,j)$, that is, the filling content of occluded area and the generating content of non-occluded area are more similar to the corresponding position of the original un-occluded expression image. Therefore, the pixels difference between the occluded and non-occluded areas between the generated image and the original un-occluded image are measured by the square of the L2 norm. The smaller the value of L2 norm is, the higher the similarity between the generated image and the original un-occluded image is. The triplet loss function l_t is a measure of the difference of the square of L2 norm between the generated image $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$, the original un-occluded image $\mathbf{Y}(i,j)$, and the occluded image $\mathbf{X}(i,j)$. The generator learns the difference of norm square between the three types of images to ensure that the generated un-occluded expression image is similar to the real un-occluded expression image in the non-occluded area.

Further, we will describe the discriminators D_o, $o = 1, 2$. Different from other studies, this study takes the de-occluded generated images, the original un-occluded images and the irrelevant-area occluded images together as the input of two discriminators. Each discriminator implements different discriminant processing under the constraints of its respective loss function, The discriminator D_1 mainly determines the real/fake of the generated images, to make the de-occluded expression image complemented by the generator more realistic, D_1 and G must form a zero-sum game, through the optimization of the adversarial loss function l_a, the filled de-occluded expression images will be more similar to the real ones. The discriminator D_2 mainly predicts the

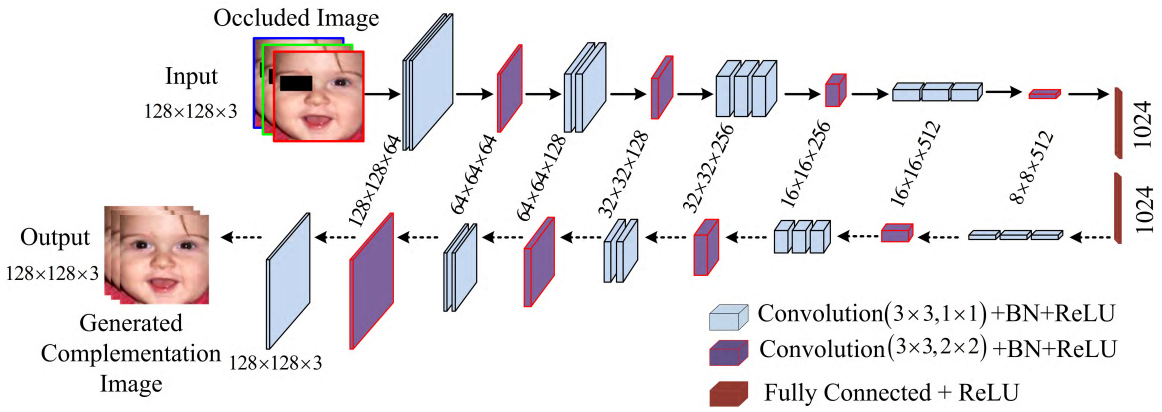


FIGURE 2. The architecture of the proposed generator.

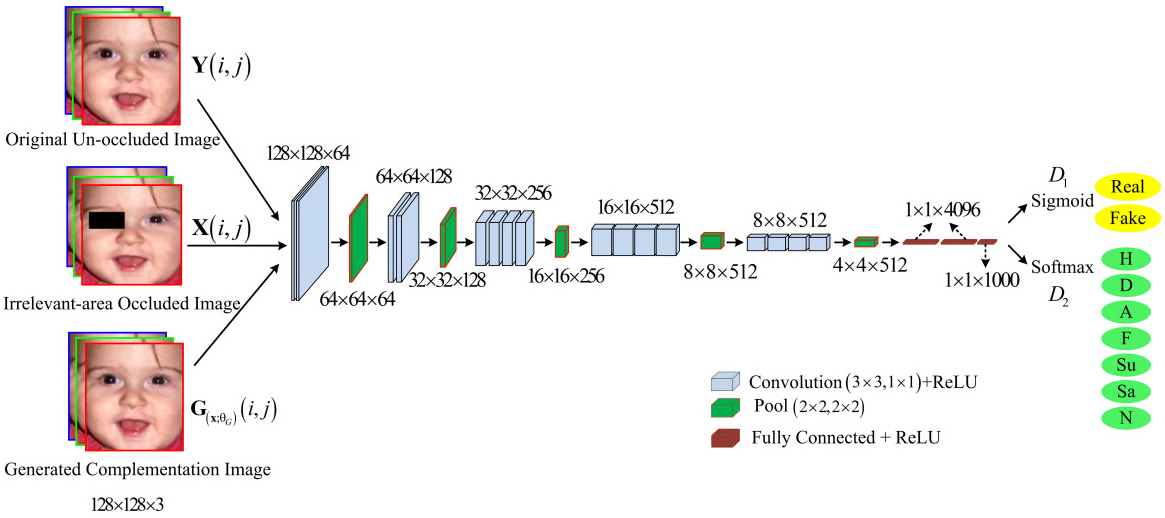


FIGURE 3. The architecture of the proposed discriminators.

classification labels (including happiness, sadness, surprise, disgust, fear, anger, and neural) of facial expression images. To recognize facial expressions, the computer needs a large number of labeled facial expression image training data, the difference between our training and the traditional training is that we take the irrelevant-area occluded images as the input of the classifier discriminator, it is mainly considered that when the occluded area is relatively small, the occluded area is in an insignificant area such as the forehead, chin, cheek, and only one eye is masked. The classifier trained by adding irrelevant area occluded images can extract more real expression features than the generated images, in this case, the separability of feature set is better than that of using only generated images and original un-occluded images. Finally, the parameters of generator and discriminators are determined by optimizing the whole network using the four loss functions constructed in this study. The structure of generator and discriminators networks is described below, and the loss function is described in the next section.

The generator constructed in this study consists of an encoder, two fully connected layers and a decoder, as shown in Fig. 2. The encoder is composed of convolution layers with output channels (64, 128, 256, 512), which uses the convolution kernel of size 3×3 and the step size of 1 to sample the occluded expression images, where the blue cube shown in Fig. 2 represents the convolution layers. Each convolution layer is followed by a Batch Normalization (BN) layer to prevent covariate shift, in addition, non-linear activation function (ReLU) is used as the activation function for each convolution layer. Then the convolution kernel of size 3×3 and the step size of 2 is responsible for down-sampling, the purple cube in Fig. 2 represents the down-sampling layers. The decoder complements de-occluded images based on the information collected by the encoder through de-convolution operation. The encoder and decoder are connected by two fully connected layers with 1024 neurons, as shown in the red-brown cube in Fig. 2.

The structures of the two discriminators constructed in this study are based on the VGG-19 network [55]. Fig. 3 shows the detailed structure of the discriminators, in which the blue cube represents the convolution layers, the green cube represents the pooling layers, and the red-brown cube represents

the fully connected layers. The two discriminators are composed of convolution layers, pooling layers, fully connected layers, and classification layer. The output channels of the convolution layers are (64, 128, 256, 512, 512), and the convolution kernel size is 3×3, and the step size is 1, the output channels of the pooling layers are (64, 128, 256, 512, 512), the filter size is 2×2, and the step size is 2, the number of neurons in the three fully connected layers is 4096, 4096, and 1000, respectively. ReLU is used as the activation function of each convolution layer and the fully connected layers. The only difference is that the discriminator D_1 finally connects the Sigmoid classification layer to complete the real and fake image binary classification under the constraint of the adversarial loss function l_a. Meanwhile, the discriminator D_2 finally connects with SoftMax classification layer, and predicts the category labels of the expression images under the constraint of the classification loss function l_c.

B. LOSS FUNCTION

1) WEIGHTED RECONSTRUCTION LOSS

As shown in Fig. 4, to make the generated de-occluded images $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$ more similar to the original un-occluded images $\mathbf{Y}(i,j)$, that is, both the complementing of the occluded area and the un-occluded area are more similar to the corresponding positions of the original un-occluded expression images, in this study, a weighted reconstruction loss function l_{wr} is introduced to measure the authenticity of the complemented images,

$$l_{wr} = \gamma_1 \delta_1 + \gamma_2 \delta_2 \tag{1}$$

It is assumed that the occluded area in the generated image $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$ is represented by O_a (i.e., the red square area shown in Fig. 4), and its size is $y \times m$. As shown in Fig. 4, the green square area represents the corresponding position of O_a in the original un-occluded image $\mathbf{Y}(i,j)$. Further, δ_1 represents the L2 norm square of the pixel value difference between $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$ and $\mathbf{Y}(i,j)$ at the corresponding position (i,j) of the occluded region, that is,

$$\delta_1 = \sum_{(i,j)\in \mathrm{O_a}} \Big[\big(\mathbf{Y}(i,j) - \mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)\big)^{\mathrm{T}} \big(\mathbf{Y}(i,j) - \mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)\big) \Big] \tag{2}$$

Similarly, δ_2 represents the L2 norm square of the pixel value difference between $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$ and $\mathbf{Y}(i,j)$ at the corresponding position (i,j) of the un-occluded area U_{sa}, that is,

$$\delta_2 = \sum_{(i,j)\in U_{\mathrm{sa}}} \Big[\big(\mathbf{Y}(i,j) - \mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)\big)^{\mathrm{T}} \big(\mathbf{Y}(i,j) - \mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)\big) \Big] \tag{3}$$

γ_1 and γ_2 in (1) denote the weights of δ_1 and δ_2, we use the ratio of occluded area and un-occluded area to the size of the whole image to determine the value of weight γ_1 and γ_2,

$$\gamma_1 = \frac{10y \times m}{128 \times 128}, \quad \gamma_2 = \frac{128 \times 128 - y \times m}{128 \times 128} \tag{4}$$

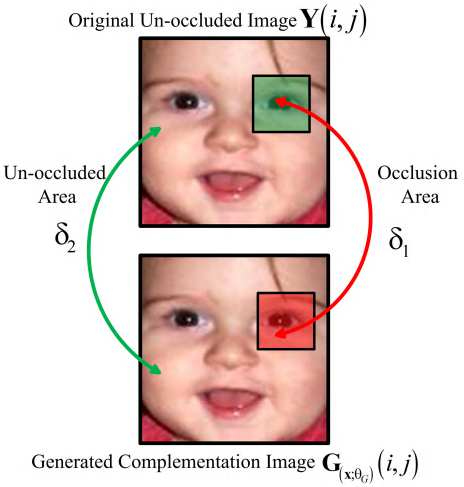


FIGURE 4. Weighted reconstruction loss function.

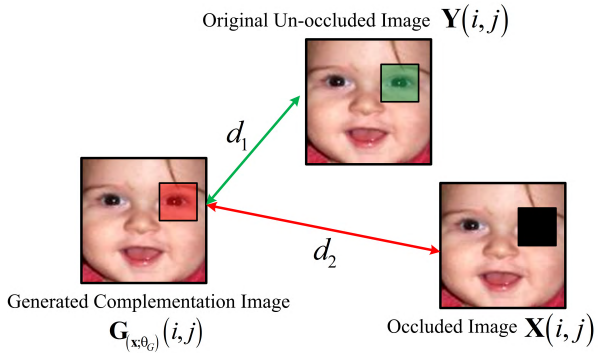


FIGURE 5. Triplet loss function.

the size of the whole image is the normalized size 128×128, the smaller the value of l_{wr} is, the higher the similarity between the generated images and the original un-occluded images in the occluded and un-occluded areas is.

2) TRIPLET LOSS

As shown in Fig. 5, the similarity between images can be represented by the spatial distance between images. In this paper, the triplet loss l_t is introduced mainly considering that l_t can better improve the details of the de-occluded images. When the two sample images are similar, the triplet loss can better model the details of the image, which is equivalent to measuring the difference between the two samples. Therefore, the network in this paper can learn better representation of the input samples, so as to optimize the details of the de-occluded images.

To ensure that the generated un-occluded area expression images $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$ are similar to the original un-occluded expression images $\mathbf{Y}(i,j)$, but different from the occluded images $\mathbf{X}(i,j)$ in the occluded area, we assume that $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$ and $\mathbf{Y}(i,j)$ are the same category samples, and $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$ and $\mathbf{X}(i,j)$ are the different categories samples. The un-occluded area of both the occluded and original un-occluded image is always the same, so we only considering

the occluded area and the corresponding area in the un-occluded and generated images in our triplet loss. As shown in Fig. 5, the black rectangle in $\mathbf{X}(i,j)$ is the occluded area, and the red and green rectangles in $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$ and $\mathbf{Y}(i,j)$ are the corresponding areas. In this study, d_1 represents the L2 norm square of the spatial distance between the red rectangle in $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$ and green rectangle in $\mathbf{Y}(i,j)$, that is the distance between the same category samples. d_2 represents the L2 norm square of the spatial distance between the black rectangle in $\mathbf{X}(i,j)$ and the red rectangle in $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$, that is the distance between the different categories samples. The ultimate optimization goal of the triplet loss is to make d_1 smaller and d_2 larger. For the spatial distance between $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$, $\mathbf{Y}(i,j)$, and $\mathbf{X}(i,j)$, there may be three cases as follows. (1) $d_1 + \beta < d_2$, the distance between $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$ and $\mathbf{Y}(i,j)$ is very close, and the distance between $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$ and $\mathbf{X}(i,j)$ is very far, at this time, there is no need to optimize. (2) $d_1 > d_2$, the distance between $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$ and $\mathbf{Y}(i,j)$ is very far. (3) $d_1 < d_2 < d_1 + \beta$, the distance between $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$ and $\mathbf{X}(i,j)$ is very close; however, there is a threshold value β between the two images.

Therefore, the triplet loss can be expressed as: $l_t = \max[(d_1 - d_2 + \beta), 0]$. To ensure that the generated un-occluded expression images are similar to the original un-occluded expression images, but different from the occluded image in the occluded area, d_1 must be very small and d_2 must be very large. In the optimization process, the network is trained by searching for the maximum value between $(d_1 - d_2 + \beta)$ and 0, that is, $\max[(d_1 - d_2 + \beta), 0]$, and then minimizing the maximum value, that is, minimizing $l_t = \max[(d_1 - d_2 + \beta), 0]$. Therefore, the ultimate triplet loss can be expressed as,

$$l_t = \min\{\max[(d_1 - d_2 + \beta), 0]\} \tag{5}$$

here, $\max[(d_1 - d_2 + \beta), 0]$ denotes the maximum of distance difference $(d_1 - d_2 + \beta)$ and 0,

$$\begin{cases} d_1 = \sum\limits_{(i,j)\in O_a} \left\|\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j) - \mathbf{Y}(i,j)\right\|^2 \\ d_2 = \sum\limits_{(i,j)\in O_a} \left\|\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j) - \mathbf{X}(i,j)\right\|^2 \end{cases} \tag{6}$$

where O_a represents the occluded area in $\mathbf{X}(i,j)$ and the corresponding areas in $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$ and $\mathbf{Y}(i,j)$, β is the threshold value.

In summary, this study introduces the triplet loss l_t and optimizes the training network by minimizing the value of $\max[(d_1 - d_2 + \beta), 0]$. This can enhance the details of the de-occluded images and to make $\mathbf{G}_{(\mathbf{x};\theta_G)}(i,j)$ closer to the distribution of $\mathbf{Y}(i,j)$ and make the generated image more realistic.

3) ADVERSARIAL LOSS

The original GAN adopts Jensen–Shannon divergence to measure the distance between probability distributions. However, when there is no intersection between the generated distribution and the real distribution, GAN will face with the instability of generator gradient, and the imbalanced punishment for diversity and accuracy will result in mode collapse [50]. However, the loss function value of WGAN [56] provides a quantitative standard for the quality of the generated image. A smaller loss value means that the generated image is more authentic. In addition, when training WGAN, it is not necessary to carefully balance the training process of generator network and discriminator network, instead, we can optimize the discriminator network until convergence, and then update the generator network to make the whole networks converge faster. Considering the above advantages, this study builds the adversarial loss function l_a based on WGAN.

WGAN determines the real/fake of the generated images through Wasserstein distance, for this study, the adversarial loss function l_a based on WGAN can be expressed as follows,

$$\begin{aligned} l_a = \min_G \max_{D_1} & (E_{y\sim p_y(y)}\left[D_1\left(y;\theta_{D_1}\right)\right] \\ & + E_{x_s\sim p_{x_s}(x_s)}\left[D_1\left(x_s;\theta_{D_1}\right)\right]) \\ & - E_{x\sim p_x(x)}\left[D_1\left(G\left(x;\theta_G\right);\theta_{D_1}\right)\right] \end{aligned} \tag{7}$$

where y represents the original un-occluded images, $p_y(y)$ is the probability distribution of y, x_s represents the discriminator's input irrelevant area occluded images, $p_{x_s}(x_s)$ is the probability distribution of x_s, x represents the generator's input occluded images, $p_x(x)$ is the probability distribution of x, $G(x;\theta_G)$ represents the de-occluded images after filling by the generator, θ_{D_1} and θ_G represent the parameters of D_1 and G, respectively, E is the expected value operator.

4) CLASSIFICATION LOSS

We takes the de-occluded generated images, original un-occluded images, and irrelevant area occluded images together as the input of classification discriminator, and we use the cross entropy loss to train the expression multi-class discriminator D_2, cross entropy loss describes the distance between the predicted probability distribution and the actual probability distribution, the classification loss function l_c in this study is expressed as follows,

$$l_c = -\sum_{f\in\mathbf{F}}\sum_{r=1}^{7} p_q(q_r)\log\left(q_r\left|f;\theta_{D_2}\right.\right) \tag{8}$$

where, θ_{D_2} represents the parameter of D_2, f represents the input prediction images, q_r , $r = 1, 2, 3, 4, 5, 6, 7$ stands for the seven real expression classification labels, $p_q(q_r)$ represents the probability distribution of q_r , $r = 1, 2, 3, 4, 5, 6, 7$, whereas $\log\left(q_r\left|f;\theta_{D_2}\right.\right)$ represents the probability distribution of prediction, that is, after the prediction of D_2, the probability that image f belongs to various expression labels. l_c is the distance between the actual output probability and the expected output probability, that is, the smaller the value of the cross entropy, the closer the two probability distributions are.

5) FULL LOSS

According to Eqs. (1,5,7), the overall objective function to optimize G can be written as:

$$L_G = \lambda_1 l_{wr} + \lambda_2 l_t + \lambda_3 l_a \tag{9}$$

Further more, according to Eqs. (7,8), the overall objective function to optimize D can be written as:

$$L_D = \lambda_3 l_a + \lambda_4 l_c \tag{10}$$

Detailed descriptions of each loss function are described in the manuscript. Where λ_1, λ_2, λ_3, λ_4, are the weight factors, which are mainly used to balance the proportion of the four loss functions."

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL SETUP

1) DATASETS

An excellent facial expression dataset is crucial to the FER system. Currently, some widely used datasets include Japanese Female Facial Expression (JAFFE) [57], Extended Cohn–Kanade (CK+) [58], AffectNet [59], and real-world affective faces database (RAF-DB) [60]. Among the above expression databases, JAFFE and CK+ have a common deficiency, that is, most of the expression images collected in these databases are frontal faces, and the expression state is extreme, and for the current mainstream deep learning algorithms, the magnitude of these databases is far from enough. In addition to high recognition rate, a good FER system should also be able to effectively recognize the natural facial expression images in real life. AffectNet and RAF-DB are selected as the experimental datasets in this study, because these two datasets reflect the natural facial expressions in real life, and the number of images is abundant.

The number of images in the AffectNet database exceeds 1 million, and the expression images it contains are all collected from the Internet. Of the one million images, about half (440K) were manually labeled in seven facial expression categories, and AffectNet is by far the world's largest database of facial expressions. In this study, we select about 190000 images with seven facial expression classification labels (anger, disgust, fear, happiness, sadness, surprise and neutral) from 440K images as training samples and select about 43000 images as testing samples. RAF-DB contains about 30000 expression images, which are collected from the Internet. Each image has about 40 independent labels, and the facial expression images in the database are different in age, gender, race, head posture, lighting conditions and other aspects, which compound the characteristics of the expression images in real life. Similar to the AffectNet dataset, RAF-DB has seven classification expression images. In this study, we select 16489 images with expression classification labels for experiments, of which 13307 are used as training samples and 3182 are used as testing samples. Table 1 shows the details of the two databases selected in this article.

TABLE 1. Detailed setup of experimental dataset in this study.

Dataset	AffectNet		RAF-DB	
Attribute	Training	Testing	Training	Testing
Happiness	90137	21088	5136	1306
Sadness	16879	4095	2471	506
Surprise	10162	2269	1301	308
Fear	4513	998	297	69
Disgust	3058	549	754	152
Anger	16545	3991	741	164
Neutral	49815	10947	2607	677
Total	191109	43937	13307	3182

2) FACIAL EXPRESSION IMAGE PREPROCESSING

To avoid the interference of illumination, posture and other factors on FER, and simultaneously, to ensure the consistency of face size, position and image quality, we first preprocess the images in the datasets, mainly including face detection, face alignment, image size and pixel normalization. The image preprocessing steps in this study are as follows. (1) Face detection. In this paper, MTCNN is used for face detection, through which the images without face can be filtered. MTCNN consists of three networks: Proposal Network (P-Net), Refine Network (R-Net), and Output Network (O-Net). Firstly, MTCNN will scale the input image in different proportions to form an image pyramid, as shown in Fig. 6(b). Secondly, we obtain the candidate windows of face region and the regression vectors of boundary boxes by P-Net, as shown in Fig. 6(c). After that, we employ Non-Maximum Suppression (NMS) to merge highly overlapped candidates, as shown in Fig. 6(d). Thirdly, R-Net filters out all candidate windows and excludes a large number of non-face windows, as shown in Fig. 6(e). The overlapping window is removed by NMS, and the result is shown in Fig. 6(f). Finally, O-Net can further filter the face candidate windows of the previous step, as shown in Fig. 6(g). In addition, O-Net can produce final bounding box and facial landmarks' positions, as shown in Fig. 6(h). (2) Face alignment. According to the physiology of human faces, the two eye centers in a face image without angle deviation is located on the same horizontal line. So, if we can get the angle θ_1 between the connecting line of the two eye centers and the horizontal line. Then, the alignment image can be obtained by rotating the image counterclockwise according to θ_1. Firstly, the human two eye centers of the facial image are located by MTCNN, and then the two eye centers are connected with a straight line segment l_1, as shown in Fig. 6(i). Secondly, through the key point of the nose, make the perpendicular line segment l_2 of l_1, as shown in Fig. 6(i). Assume the coordinates of the left eye are (x_1, y_1), and the coordinates of the right eye are (x_2, y_2). When the face in the image have the angular deviation, the angle formed by l_1 and horizontal axis l_3 is θ_1, while the angle formed by l_2 and horizontal axis l_3 is θ_2. We take the key point of right eye as the center to rotate the face image counterclockwise by $\theta_1 = \arctan\left|(y_1 - y_2)/(x_1 - x_2)\right|$ degrees, which is to complete the face alignment. At this time

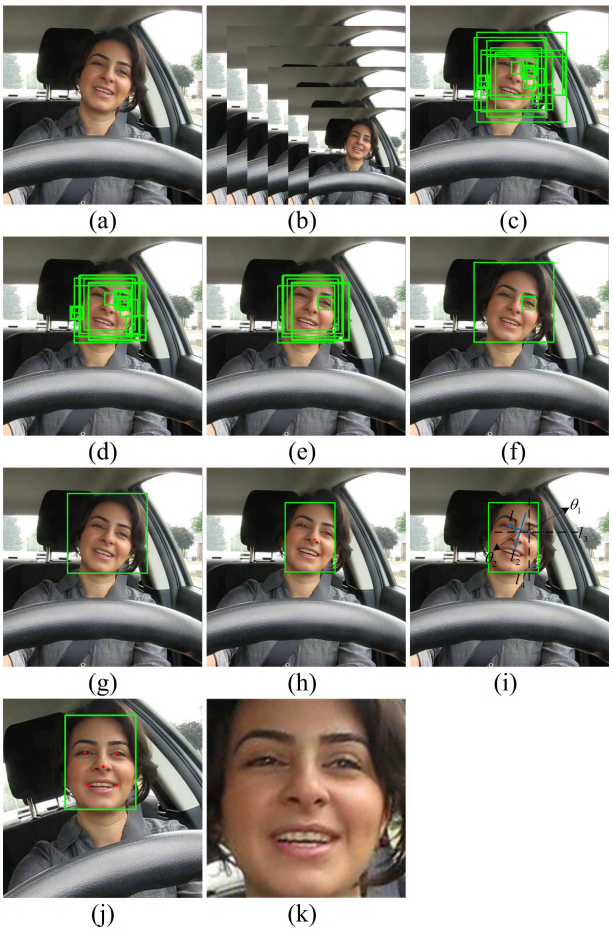


FIGURE 6. Expression image preprocessing. (a) Original. (b) Image pyramid. (c) P-Net processing. (d) Bounding box regression of P-Net. (e) R-Net processing. (f) Bounding box regression of R-Net. (g) O-Net processing. (h) Face detection result. (i) Face calibration. (j) Face calibration result. (k) Final image.

$\theta_2 = 90°$, face alignment is completed. The face image calibrated by our method is shown in Fig. 6(j). (3) Normalization is performed to obtain standardized face images with the same size and the same range of gray values. After face detection and alignment processing, we crop the face region from the original facial image according to the green bounding box, as shown in Fig. 6(j). The original bounding box is not always the square, but we cropped the face region according to the bounding box, and the fixed shape is set to square. At last, we use the scaling transformation to unify all the cropped images into the fixed size 128×128. Meanwhile, the image pixel values are normalized from range [0, 255] to range [0, 1]. The final preprocessing result is shown in Fig. 6(k).

3) OCCLUDED IMAGE SIMULATION

For un-occluded FER, there are mature and access an experimental datasets, but there is almost no mature and standard dataset for researchers to carry out occluded FER. Therefore, we decide to use the images in AffectNet and RAF-DB database as the prototype to simulate occlusion. In real life, there are generally two types of occlusions, one is temporary occlusion caused by hand or head movement, and the other is systematic occlusion caused by sunglasses, masks and scarves. Therefore, this study simulates the possible occlusion in reality by adding black rectangles of different sizes at different positions of the facial expression images.

FIGURE 7. Occlusion simulation examples. (a) Left eye. (b) Right eye. (c) Two eyes. (d) Left half face. (e) Right half face. (f) Mouth. (g) Nose. (h) Random 20%. (i) Random 30%. (j) Random 40%.

TABLE 2. Detailed setup of occluded experimental dataset in this study.

Database	AffectNet		RAF-DB	
Occlusion type	Training	Testing	Training	Testing
Left eye	7761	1439	1174	229
Right eye	7673	1350	986	181
Two eyes	6607	1308	821	153
Left half	5320	1027	737	132
Right half	5807	1178	771	128
Mouth	5627	949	603	149
Nose	5432	986	636	137
Random 20%	2009	397	341	76
Random 30%	1104	221	318	60
Random 40%	561	87	262	49
Total	47901	8942	6649	1294

We set a total of 10 occlusion situations, as shown in Figs. 7(a)–7(j), in addition to the eyes, mouth and nose, the random occluded area of 20%, 30%, and 40% are also considered. By observing Fig. 7(j), it can be seen that when the occluded area exceeds 40%, all facial features have been covered, at this time, it is of little significance to filling the occlusion, therefore, it is not necessary to analyze the case that the occluded area is larger than 40%. The details of the occluded dataset after the simulated occlusion are shown in Table 2.

4) IMPLEMENTATION DETAILS

Our experiments are conducted on a GPU workstation by NVIDIA GeForce GTX-1080Ti 12G [61]. For training,

the initial learning rate was 0.01, the batch size was set to 48, and the iterations as 20k. In the actual training process, we first train the discriminators, and then train the generator. The discriminators and the generator are trained alternately, in each round, we first train the discriminators once, and then train the generator 5 times. In this paper, discriminators D_1 and D_2 are trained simultaneously.

To balance the influence of different loss functions, we set $\lambda_1 = 100$, $\lambda_2 = 10$, $\lambda_3 = 1$, and $\lambda_4 = 1$ in the experiment.

B. QUALITATIVE EXPERIMENTAL RESULTS AND DISCUSSION

1) VISUALIZATION RESULT OF DE-OCCLUSION

In order to show the complemented of the generator under different loss function optimizations, first, the generator is optimized under the triplet loss l_t, the adversarial loss l_a, and the weighted reconstruction loss l_{wr}, respectively. The generated images are shown in Figs. 8(c), 8(d), and 8(e). It can be seen that none of the loss functions is good enough by itself. Meanwhile, when a single loss function is used to optimize the generator, the generated image is not good. Compared with l_t and l_a, although the generated image under the optimization of l_{wr} is fuzzy and smooth, we can still clearly see the organs contour of the face and it is the most similar image to the original unoccluded image. Then, the generator is optimized under $l_{wr}+l_a$ and $l_{wr}+l_t$, respectively, as generated images Figs. 8(f) and 8(g). It can be seen that Fig. 8(f) is more realistic than Fig. 8(e), and the occluded area in Fig. 8(g) is more detailed than Fig. 8(e). However, when only combined l_{wr} with another loss function (l_t or l_a), some details of the generated image cannot be complemented or the image is not real enough. In order to get the more real and clearer face images, we optimized the generator with $l_{wr} + l_t + l_a$, the generated image is shown in Fig. 8(h). The results show that the three loss functions have different effects, and the good generated image can be obtained only when three loss functions are combined. The results also show that none of the loss functions is good enough by itself and all of them are indeed needed for image complementing.

To show that the proposed method can effectively complement the occluded facial expression images, the partial de-occlusion visualization results are shown in Fig. 9. Figs. 9(a) and 9(b) show the de-occluded of the database RAF-DB and AffectNet under different occlusion conditions, respectively. The three images in each row of the two subgraphs (a) and (b) in Fig. 9 are a group, in which the first one is the original un-occluded expression image, the second one is the occluded simulation image, and the third one is the generated de-occluded image. In each column, from top to bottom, there are nose occlusion, single-eye occlusion, two-eyes occlusion, mouth occlusion, unilateral face occlusion, random occlusion 20%, random occlusion 30%, and random occlusion 40%.

It can be seen from the first and second lines in Figs. 9(a) and 9(b) that the results of complementing the nose and the single-eye occluded area is good, and our method can fill the occluded area naturally. The complemented images cannot be completely same with the original images; however, the slight difference does not change the classifications of the expression images and does not affect the extraction of expression features and expression recognition. It can be seen from the third and fourth lines in Figs. 9(a) and 9(b) that there is a gap between the complementing de-occluded results of two-eyes (mouth) and the original images. We analyze that because the eye is the window to the soul, it can most directly and completely show the mental state and inner activities of a person. Compared with the occlusion of a single eye, there is no reference when both eyes are occluded at the same time. Therefore, there are differences in shape and pupil color between the

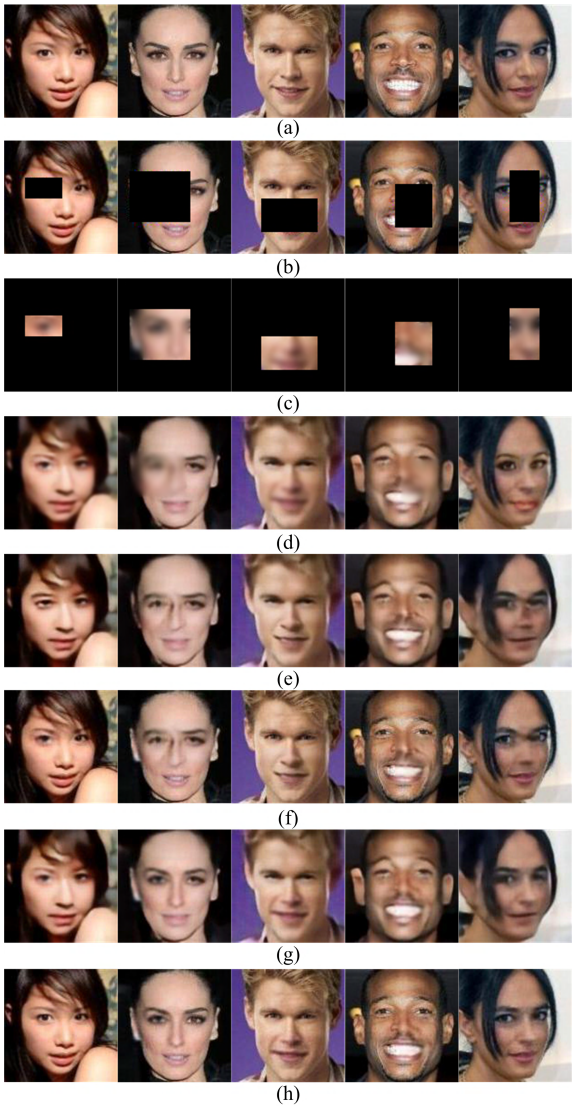


FIGURE 8. Visualization results of complementing occluded area under different loss function optimizations. (a) Original un-occluded expression image. (b) Occluded simulation image. (c) Generated image under the effect of l_t. (d) Generated image under the effect of l_a. (e) Generated image under the effect of l_{wr}. (f) Generated image under the effect of $l_{wr} + l_a$. (g) Generated image under the effect of $l_{wr} + l_t$. (h) Generated image under the effect of proposed loss ($l_{wr} + l_t + l_a$).

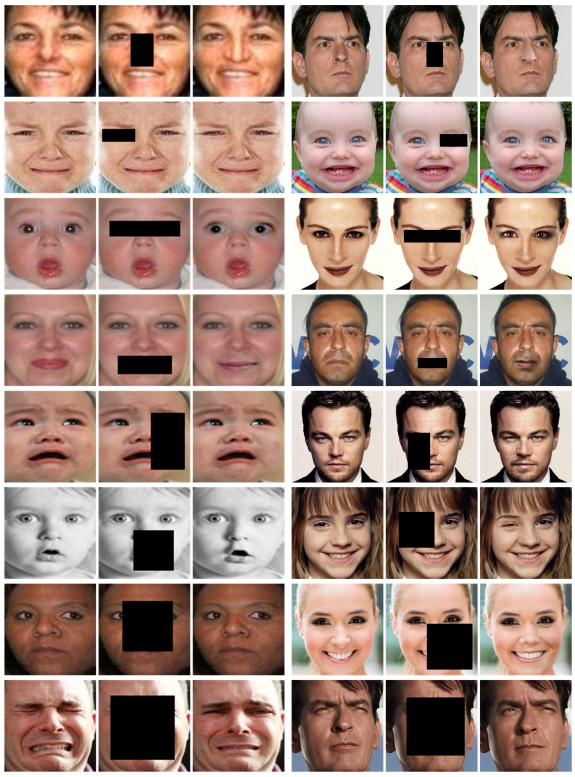


FIGURE 9. Visualization results of complementing occluded area on the testing datasets (a) RAF-DB and (b) AffetNet.

filled eyes and the original images, but the classification of generated de-occluded expression images does not change. The expression of the mouth is mainly reflected by the change of the mouth shape. By observing the fourth line in Figs. 9(a) and 9(b), it can be observed that the appearance of the generated filling mouth is slightly different from the previous one; however, the mouth shape does not change significantly, and the expression classification of the whole image remains the same. Therefore, it can be proved that the weighted reconstruction loss and triplet loss proposed in this study can effectively shorten the spatial distance between the occluded area and the original un-occluded image. It can be seen from the fifth and sixth lines in Figs. 9(a) and 9(b), in the case of a unilateral face and a random occlusion of 20%, the occluded areas after the filling have some changes in eye size and mouth contour as compared to the original un-occluded images. It can be seen that, compared with the previous occlusion situation, the increase of the occluded area and the randomness do increase the difficulty for the filling work; however, the generated de-occluded images still have the same expression classification as the original images, and still does not affect the subsequent expression recognition. It can be seen from the seventh and eighth lines in Figs. 9(a) and 9(b), as the random occluded area increases, the gap between the generated de-occluded images and the original images becomes larger. More obvious is the last row in Figs. 9(a) and 9(b), a large range of partial occlusion may causes loss of most of the facial features; thus, increasing the difficulty of de-occlusion. As the occluded area increases, the generated complemented may ''create'' something different from the original image. Therefore, when the occlusion area is larger than 40%, we believe that the generated de-occluded image has a large gap with the original un-occluded image and is no longer suitable for subsequent expression recognition.

In summary, the weighted reconstruction loss proposed in this study gives different weights to the occluded area and the un-occluded area in the occluded image. Moreover, we introduce the triplet loss between the original un-occluded image, occluded image, and generated de-occluded image, so that it can effectively fill single-eye, two-eyes, mouth, half face, and irrelevant area occlusion; and the result of filling is natural and similar to the original un-occluded image content.

2) VISUALIZATION RESULT OF DE-OCCLUSION FOR REAL OCCLUDED IMAGES

By observing the existing research of occluded facial expression recognition, we found that there is no universal, mature and data-rich occluded facial expression database at present. Moreover, the real occlusion images are much more difficult to handle, because it is not easy to detect as occlusion or obtain training images with both unoccluded/occluded versions. Therefore, at present, occluded expression recognition is mostly carried out on the occluded simulation database established by researchers themselves. Most researchers used mask images (e.g. black rectangles, apples, cups, glasses, and so on) to splice with the unoccluded expression images to synthesize simulated occluded data.

In order to verify the de-occluded of our method for real occluded images, we download several real occluded facial expression images from the Internet for testing. We tested three real occluded facial expression images, e.g. sunglasses, hands, and tape sticking to the mouth. The experimental results are shown in Fig. 10, where Fig. 10(a) show the real occluded images and Fig. 10(b) show the de-occluded of our method. It can be seen from Fig. 10(b) that the de-occluded of the proposed method is not ideal when dealing with the real occluded expression images, and there is still much room for improvement. This is because this study simulated the possible occlusion in reality by adding black rectangles of different sizes at different positions of the facial expression images, and we use simulation data as training dataset. The network in this paper lacks sufficient real occluded facial expression images as training data. Therefore, when testing the real occluded images, our network can not effectively complement the area where shelters exist, such as sunglasses, hands, and tape.

For the existing occluded facial expression recognition, no matter what kind of mask images the researchers use to simulate the occluded images, the simulation results are far from the real occluded facial expression images. No methods can effectively process the realistic occluded facial expression images without sufficient real occlusion training data.

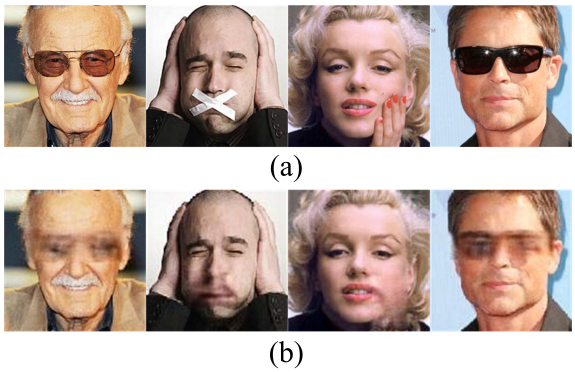


FIGURE 10. De-occluded of real occluded facial expression images. (a) Realistic occluded facial expression images. (b) The de-occluded of our method.

3) 3D T-SNE FEATURE MAP

The difference between this study and the traditional training method is that the irrelevant area occluded images, the original un-occluded images, and the generated de-occluded images are taken all together as the input of the classifier discriminator. It is mainly considered that facial expression is determined by eyebrows, eyes, nose, and mouth, it has a weak relationship with forehead, chin and cheek. For this reason, we named the images whose occlusion position is located in the forehead, chin, and cheek as the irrelevant area occluded images. The classifier trained by adding irrelevant area occluded images can extract more real expression features than the generated images, In this case, the separability of feature set is better than that of using only generated images and original un-occluded images. To show that the proposed training method can help the discriminator extract more separable facial features, the facial features are displayed with T-distributed Stochastic Neighbor Embedding (t-SNE) [62] visualization results. Taking 1300 random samples in RAF-DB dataset as an example, the feature distributions under the two training methods are shown in Figs. 11(a) and 11(b). Figure 11(a) represents the proposed training with the original unoccluded, the generated unoccluded, and the irrelevant area occluded images. Figure 11(b) represents training with only original and the generated unoccluded images. In the legend in Fig. 11, H represents happiness, N represents Neutral, A represents anger, Sa represents sadness, D represents disgust, Su represents surprise, and F represents fear. It can be seen that, compared with Fig. 11(b), the proposed training shown in Fig. 11(a) classifies the features of seven types of expressions more clearly, which provides a powerful help for subsequent classification and recognition. Therefore, the training method proposed in this study is effective.

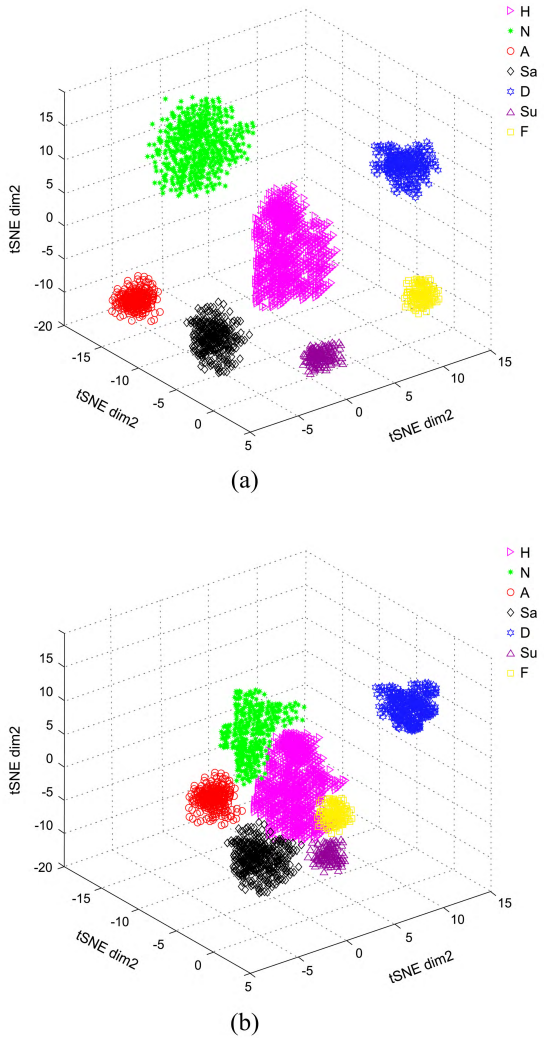


FIGURE 11. 3D t-SNE feature visualization results of RAF-DB training dataset under different training methods. (a) The proposed training with original un-occluded, generated un-occluded, and irrelevant area occluded images. (b) Training with only original and generated un-occluded images.

C. QUANTITATIVE EXPERIMENTAL RESULTS AND DISCUSSION

To analyze the effect of face complementation on occluded FER, this study conducts a comparison experiment on the recognition rate of facial expression images before and after occlusion area filling on the RAF-DB dataset. For each type of occlusion, the number of expression images before and after the filling is the same, we put the facial expression images before and after filling into the recognition network for expression discrimination, and the experimental results are shown in Fig. 12. As shown in Fig. 12, the purple regions represent the recognition rate of the generated un-occluded images, and the pink region represents the recognition rate of the occluded images. For the image without occlusion, the filling process has no effect and will not affect the quality of the image, and the recognition rate of the two is the same. However, with the addition of occlusion area, especially when occlusion is located in eyes, mouth, and so on, the recognition rate of the images after occlusion filling processing is improved significantly as compared with the occluded images, especially with the expansion of the occluded area and the particularity of occluded position, the recognition rate

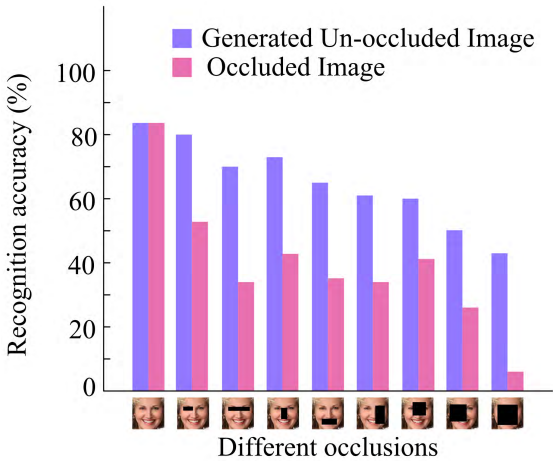


FIGURE 12. Comparison of recognition rates before and after de-occlusion in the RAF-DB dataset.

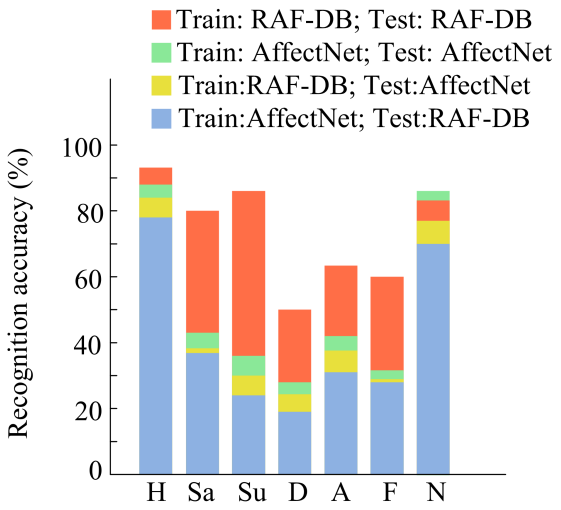


FIGURE 13. Recognition rates of occluded facial expression images under different training/testing datasets.

of the occluded images decrease sharply; however, the recognition rate of de-occluded images still remain above 45%.

Overall, the de-occlusion proposed in this study improves the recognition rate of the occluded expression images, and this method is significant for the recognition of occluded expression images. The de-occlusion has complemented the key areas of expression features. Therefore, the facial features in the generated de-occluded images are more abundant and accurate, which can provide efficient help for subsequent classification and recognition.

Furthermore, to verify that the method proposed in this study is independent of database and has strong robustness, we used the AffecNet/RAF-DB dataset for training and then testing our method on the RAF-DB/AffecNet dataset. We validate the robustness of our algorithm on the simulated occluded dataset established in this paper. The details of the occluded dataset are shown in Table 2, and the number of images of each category of expression in the occluded dataset is shown in Table 3. Four experiments are carried out on the WGAN-based occluded facial expression recognition network proposed in this paper. (1) Training network with RAF-DB training dataset and testing with RAF-DB testing dataset. (2) Training network with AffectNet training dataset and testing with AffectNet testing dataset. (3) Training network with RAF-DB training dataset and testing with AffetNet testing dataset. (4) Training network with AffectNet training dataset and testing with RAF-DB testing dataset. The average recognition rates of seven categories of occluded facial expressions under four experimental conditions are shown in Fig. 13. In the legend in Fig. 13, H represents happiness, Sa represents sadness, Su represents surprise, D represents disgust, A represents anger, F represents fear, and N represents Neutral. The experimental results show that the proposed method has a high recognition rate for a single database. In addition, when training with one database and testing on another database, the expression recognition rate is also satisfactory, which shows that the algorithm proposed in this paper has strong robustness.

TABLE 3. Expression classification details of the occlusion database.

Database	AffectNet		RAF-DB	
Attribute	Training	Testing	Training	Testing
Happiness	22167	4196	2600	503
Sadness	3759	745	976	190
Surprise	2764	503	507	99
Fear	1125	209	514	102
Disgust	754	151	501	95
Anger	4223	746	492	89
Neutral	13109	2392	1059	216
Total	47901	8942	6649	1294

In this study, different methods are used to identify unoccluded and occluded expression images in the RAF-DB and AffectNet datasets, respectively, and the recognition rates under different methods are presented in Table 4 for comparison. We compare our method with the methods proposed in Ref. [55], [63]–[65], and [66].

In order to make the comparison fair, we have implemented these methods. To be more specific, the unoccluded/occluded results of Refs. [55], [63]–[65] are produced by ourselves. For Ref. [66], the recognition rate of the occluded facial expression is produced by ourselves, and the recognition rate of the unoccluded facial expression is the result published in this paper. We have attempted to reproduce the methods as described in [55], [63]–[65], and all the settings referred to in these papers have been implemented. However, since we have failed to contact with the authors when asking some details and network parameters which are not provided in the paper, these details and parameters are set according to our experiences. Nevertheless, we also obtained the similar results. For Ref. [66], our result is slightly different from the recognition rate published in this paper for unoccluded facial expression recognition, so we reference its published result. But the recognition rate of occluded expression is produced by ourselves. All the methods in Table 4 are implemented

TABLE 4. Average recognition rate among different methods on AffectNet and RAF_DB datasets.

Methods	RAF-DB		AffectNet	
Attribute	Original	Occluded	Original	Occluded
FRR-CNN[63]	79.26	74.45	50.65	45.77
VGG16[55]	80.96	75.26	51.11	46.48
CapsNet[64]	77.48	73.71	51.26	47.36
DL-CNN[65]	76.09	71.44	51.32	46.15
EAU-Net[66]	81.83	77.01	58.91	50.04
Proposed	83.49	78.35	59.73	51.21

on a GeForce GTX 1080 Ti 12G experimental platform. Like our method, the initial learning rate of all methods is 0.01, and take 20k iterations with the batch size of 48. For Ref. [63] and [65], more detailed parameter are set as follows, the learning rate was reduced by polynomial policy with gamma of 0.1, the momentum is 0.9, and the weight decay is 0.0005. All of the methods in Table 4 were trained from the beginning, and they were not training start from a pretrained network. They were also not tested on a pre-training model that had been trained on another dataset. The training and testing datasets of the methods in Refs. [55], [63]–[66] are the same as our methods. That is to say, all the occluded training datasets in Table 2 are used for training, and we also selected a small number of un-occluded facial expression images from the two datasets in Table 1, and add them together with the training datasets in Table 2 as the total training data. The reason for adding the un-occluded expression images to the training dataset is that, for the unknown input image, we do not know whether it is occluded or not in advance. In order to make our method can also achieve accurate recognition for unoccluded facial expression images, we added a small number of un-occluded facial expression images to the training dataset. Meanwhile, the occluded and irrelevant area occluded images are used for training also when the performance on original unoccluded images is evaluated.

For the RAF-DB dataset, the method proposed in this study is 4.23%, 2.53%, 6.01%, 7.40%, and 1.66% higher in the recognition rate of un-occluded facial expression images than those of the methods discussed in Ref. [55], [63]–[65], and [66], respectively. However, in terms of occluded FER, it is 3.90%, 3.09%, 4.64%, 6.91%, and 1.34% higher than them, respectively. For the AffectNet dataset, the method proposed in this study is 9.08%, 8.62%, 8.47%, 5.41%, and 0.82% higher in the recognition rate of un-occluded facial expression images than those of the methods discussed in Ref. [55], [63]–[65], and [66], respectively. In terms of occluded FER, it is 5.44%, 4.73%, 3.85%, 5.06%, and 1.17% higher than them, respectively.

By observing the research results of non-occluded facial expression recognition in the past three years, we found that the average accuracy of facial expression recognition in AffectNet and RAF-DB databases increased by about 1.197% and 1.53% respectively compared with the previous year. We also found that the rate of improvement in the accuracy of facial expression recognition is slowing down year by year. For occluded facial expression recognition, the difficulty brought by the randomness of occluded area size and location leads to a decrease in the improvement of the average accuracy of occluded facial expression recognition. Although the performance improvement of our proposed method is limited, it is basically in line with the average growth rate. The improvement of occluded expression recognition rate in this paper is mainly affected by the following two reasons. First, because the occlusion training datasets mostly consists of the single eye, double eyes, mouth, nose, and random 20% occluded situations, the comparative methods can also recognize the occluded facial expression images to some extent. It is believed that the performance of our method will be significantly better than these methods on the larger occluded area facial expression images. Second, in facial expression recognition, neutral and happiness expressions are easy to recognize, while disgust and anger expressions are easy to confuse. Because the number of happy and neutral expression images is the majority in the original unoccluded expression datasets, and the number of disgust and anger expression images is less than that, this also leads to the majority of happy and neutral occluded facial expression images in our established occluded datasets. Therefore, for the facial expression images which are easy to recognize, the difference of the recognition rate of each method will be reduced. It is believed that our method will be superior to these methods in the database with average number of seven kinds of facial expression images.

The experimental results show that the method proposed in this study is superior to the existing methods for recognition of both un-occluded and occluded expression images in the RAF-DB and AffectNet datasets. For occluded FER, the method proposed in this study can effectively fill the occluded areas and accurately classify and identify the generated de-occluded expression images.

V. CONCLUSION

Facial expressions contain rich personal emotional information, and the automatic recognition of expressions has broad application prospects in the fields of human-computer interaction, intelligent security, psychological analysis, and so on. Currently, most FER methods consider the frontal and un-occluded facial expression images as the research objects. However, in real life, expression occlusion occurs often, and occluded areas cause decrease in the recognition rate and robustness of the recognition method. Therefore, a more robust FER method for facial expression images under local occlusion conditions has become a research hotspot in the field of computer intelligence applications.

In this study, with the aim of achieving robust expression recognition under local occlusion, we propose an occluded FER method based on WGAN. This method consists of one

generator and two discriminators. The generator network mainly complements the occluded area of the facial expression images under the double constraint of the weighted reconstruction and triplet loss functions proposed in this study. In addition, on the basis of the original un-occluded images and the generated complementing images, a irrelevant area occluded images are added as the input of the discriminators. We use the Wasserstein distance to construct the adversarial loss function between the generated complemented, i.e., the original un-occluded images and the irrelevant-area occluded images, so as to optimize the feature extraction ability of the system. Finally, we introduce the classification loss function to complete the expression recognition process. In this study, a series of experiments are conducted on two in-the-wild datasets, AffectNet and RAF-DB. The experimental results show that the proposed method has satisfactory complementing effect for monocular occlusion, binocular occlusion, mouth occlusion, nose occlusion, half face occlusion, and random occlusion area are less than 40%. In addition, the expression recognition rates of the images with or without occlusion are improved much more than the state-of-the-art methods.

In the future, we will continue to study how to build a more effective generation system to truly complement and recognize the occluded facial expression images when the occlusion area is more than 40%. In addition, we will also study the problem of de-occluded and recognition of realistic occluded facial images from the following three aspects. (1) Further study and find the appropriate loss functions. (2) Find the relationship between the real occluders and the unoccluded area of the face to optimize the details of the filling area. (3) Establish an open, mature, and data-rich realistic occluded facial expression database through our own efforts. We believe that in the future, we will have new research progress on the realist occluded facial expression recognition.

REFERENCES

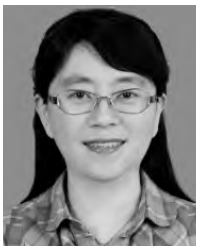
[1] Y. Ding, Q. Zhao, B. Li, and X. Yuan, "Facial expression recognition from image sequence based on LBP and Taylor expansion," *IEEE Access*, vol. 5, pp. 19409–19419, 2017.
[2] C. Qi, M. Li, Q. Wang, H. Zhang, J. Xing, Z. Gao, and H. Zhang, "Facial expressions recognition based on cognition and mapped binary patterns," *IEEE Access*, vol. 6, pp. 18795–18803, Mar. 2018.
[3] J. Jang, H. Cho, J. Kim, J. Lee, and S. Yang, "Facial attribute recognition by recurrent learning with visual fixation," *IEEE Trans. Cybern.*, vol. 49, no. 2, pp. 616–625, Feb. 2019.
[4] B.-F. Wu and C.-H. Lin, "Adaptive feature mapping for customizing deep learning based facial expression recognition model," *IEEE Access*, vol. 6, pp. 12451–12461, Feb. 2018.
[5] L. Zhao, Z. Wang, X. Wang, and Q. Liu, "Driver drowsiness detection using facial dynamic fusion information and a DBN," *IET Intell. Transp. Syst.*, vol. 12, no. 2, pp. 127–133, Mar. 2018.
[6] M. I. Chacon-Murguia and C. Prieto-Resendiz, "Detecting driver drowsiness: A survey of system designs and technology," *IEEE Consum. Electron. Mag.*, vol. 4, no. 4, pp. 107–119, Oct. 2015.
[7] Y. Chen, T. Wang, H. Wu, and Y. Wang, "A fast and accurate multi-model facial expression recognition method for affective intelligent robots," in *Proc. IEEE Int. Conf. Int. Saf. Robot. (ISR)*, Shenyang, China, Aug. 2018, pp. 319–324.
[8] I. Cohen, R. Looije, and M. A. Neerincx, "Child's perception of robot's emotions: Effects of platform, context and experience," *Int. J. Soc. Robot.*, vol. 6, no. 4, pp. 507–518, Nov. 2014.
[9] L. Zhang, M. Jiang, D. Farid, and M. A. Hossain, "Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot," *Expert Syst. Appl.*, vol. 40, no. 13, pp. 5160–5168, Oct. 2013.
[10] H. Mahersia and K. Hamrouni, "Using multiple steerable filters and Bayesian regularization for facial expression recognition," *Eng. Appl. Artif. Intell.*, vol. 38, pp. 190–202, Feb. 2015.
[11] B. Abboud, F. Davoine, and M. Dang, "Facial expression recognition and synthesis based on an appearance model," *Signal Process.-Image Commun.*, vol. 19, no. 8, pp. 723–740, Sep. 2004.
[12] F. Ren and Z. Huang, "Facial expression recognition based on AAM–SIFT and adaptive regional weighting," *Proc. IEEJ*, vol. 10, no. 6, pp. 713–722, Nov. 2015.
[13] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3D facial expression recognition: A comprehensive survey," *Image Vis. Comput.*, vol. 30, no. 10, pp. 683–697, Oct. 2012.
[14] G. Muhammad, M. Alsulaiman, S. U. Amin, A. Ghoneim, and M. F. Alhamid, "A facial-expression monitoring system for improved healthcare in smart cities," *IEEE Access*, vol. 5, pp. 10871–10881, 2017.
[15] H. Yan, "Collaborative discriminative multi-metric learning for facial expression recognition in video," *Pattern Recognit.*, vol. 75, pp. 33–40, Mar. 2018.
[16] C.-T. Liao, H.-J. Chuang, C.-H. Duan, and S.-H. Lai, "Learning spatial weighting for facial expression analysis via constrained quadratic programming," *Pattern Recognit.*, vol. 46, no. 11, pp. 3103–3116, Nov. 2013.
[17] J. Yu and Z. Wang, "A video-based facial motion tracking and expression recognition system," *Multimedia Tools Appl.*, vol. 76, no. 13, pp. 14653–14672, Jul. 2017.
[18] H. Dibeklioglu, A. A. Salah, and T. Gevers, "A statistical method for 2-D facial landmarking," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 844–858, Feb. 2012.
[19] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2018.
[20] L. Zhang, K. Mistry, M. Jiang, S. C. Neoh, and M. A. Hossain, "Adaptive facial point detection and emotion recognition for a humanoid robot," *Comput. Vis. Image Understand.*, vol. 140, pp. 93–114, Nov. 2015.
[21] Y. Wu and Q. Ji, "Feature extraction trends for intelligent facial expression recognition: A survey," *Inf. J. Comput. Inform.*, vol. 42, no. 4, pp. 507–514, Dec. 2018.
[22] J. Jin, B. Xu, X. Liu, Y. Wang, L. Cao, L. Han, B. Zhou, and M. Li, "A face detection and location method based on feature binding," *Signal Process.-Image Commun.*, vol. 36, pp. 179–189, Aug. 2015.
[23] Y. Luo and Y.-P. Guan, "Adaptive skin detection using face location and facial structure estimation," *IET Comput. Vis.*, vol. 11, no. 7, pp. 550–559, Oct. 2017.
[24] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[25] M. Zhou, H. Lin, S. S. Young, and J. Yu, "Hybrid sensing face detection and registration for low-light and unconstrained conditions," *Appl. Opt.*, vol. 57, no. 1, pp. 69–78, Jan. 2018.
[26] M. Z. Uddin, M. M. Hassan, A. Almogren, A. Alamri, M. Alrubaian, and G. Fortino, "Facial expression recognition utilizing local direction-based robust features and deep belief network," *IEEE Access*, vol. 5, pp. 4525–4536, 2017.
[27] Z. Xiang, H. Tan, and W. Ye, "The excellent properties of a dense grid-based HOG feature on face recognition compared to Gabor and LBP," *IEEE Access*, vol. 6, pp. 29306–29319, Mar. 2018.
[28] M. Z. Uddin, W. J. Khaksar, and J. Torresen, "Facial expression recognition using salient features and convolutional neural network," *IEEE Access*, vol. 5, pp. 26146–26161, 2017.
[29] T. T. D. Pham, S. Kim, Y. Lu, S.-W. Jung, and C.-S. Won, "Facial action units-based image retrieval for facial expression recognition," *IEEE Access*, vol. 7, pp. 5200–5207, 2019.
[30] Q. Jia, X. Guo, H. Guo, Z. Luo, and Y. Wang, "Multi-layer sparse representation for weighted LBP-patches based facial expression recognition," *Sensors*, vol. 17, no. 12, pp. 6719–6739, Mar. 2015.
[31] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Discriminative shared Gaussian processes for multiview and view-invariant facial expression recognition," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 189–204, Jan. 2015.
[32] Q. Wang, H. Fan, L. Zhu, and Y. Tang, "Deeply supervised face completion with multi-context generative adversarial network," *IEEE Signal Process. Lett.*, vol. 26, no. 3, pp. 400–404, Mar. 2019.

[33] S. F. Cotter, "Recognition of occluded facial expressions using a fusion of localized sparse representation classifiers," in *Proc. IEEE Digit. Signal Workshop*, Sedona, AZ, USA, Jan. 2011, pp. 437–442.
[34] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou, "UV-GAN: Adversarial facial UV map completion for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7093–7102.
[35] Y. Deng, Q. Dai, and Z. Zhang, "Graph Laplace for occluded face completion and recognition," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2329–2338, Aug. 2011.
[36] H. Wu, Z. Miao, Y. Wang, J. Chen, C. Ma, and T. Zhou, "Image completion with multi-image based on entropy reduction," *Neurocomputing*, vol. 159, pp. 157–171, Jul. 2015.
[37] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2016, pp. 2536–2544.
[38] I. Satoshi, S.-S. Edgar, and I. Hiroshi, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, Jul. 2017, Art. no. 107.
[39] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5505–5514.
[40] D. Rim, S. Honari, M. K. Hasan, and C. J. Pal, "Improving facial analysis and performance driven animation through disentangling identity and expression," *Image Vis. Comput.*, vol. 52, pp. 125–140, Aug. 2016.
[41] D. Ghimire, S. Jeong, J. Lee, and S. H. Park, "Facial expression recognition based on local region specific features and support vector machines," *Multimed. Tools Appl.*, vol. 76, no. 6, pp. 7803–7821, Mar. 2017.
[42] O. Rudovic, M. Pantic, and I. Patras, "Coupled Gaussian processes for pose-invariant facial expression recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1357–1369, Jun. 2013.
[43] R. A. Patil, V. Sahula, and A. S. Mandal, "Features classification using geometrical deformation feature vector of support vector machine and active appearance algorithm for automatic facial expression recognition," *Mach. Vis. Appl.*, vol. 25, no. 3, pp. 747–761, Apr. 2014.
[44] R. Singh, S. Jadhav, and S. K. Gupta, "Average Gabor-wavelet filter feature extraction technique for facial expression recognition," *Signal Process.*, vol. 117, pp. 10–71, Dec. 2015.
[45] R. Singh, S. Jadhav, and S. Sharma, "Facial expression recognition based on improved local binary pattern and class-regularized locality preserving projection," *Proc. IJNSA*, vol. 15, no. 5, pp. 102–104, May 2015.
[46] Y. Lu, S. Wang, W. Zhao, Y. Zhao, and J. Wei, "A novel approach of facial expression recognition based on shearlet transform," in *Proc. 5th IEEE Global Conf. Signal Inf.*, Montreal, QC, Canada, Nov. 2017, pp. 398–402.
[47] Y. Lu, S. Wang, and W. Zhao, "Facial expression recognition based on discrete separable shearlet transform and feature selection," *Algorithms*, vol. 12, no. 1, pp. 1–14, Jan. 2019.
[48] M. Melek, A. Khattab, and M. F. Abu-Elyazeed, "Fast matching pursuit for sparse representation-based face recognition," *IET Image Process.*, vol. 12, no. 10, pp. 1807–1814, Oct. 2018.
[49] T. H. Kim, C. Yu, and S. W. Lee, "Facial expression recognition using feature additive pooling and progressive fine-tuning of CNN," *Electron. Lett.*, vol. 54, no. 23, pp. 1326–1328, Nov. 2018.
[50] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, 2014, pp. 2672–2680.
[51] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," Jan. 2016, *arXiv:1511.06434*. [Online]. Available: https://arxiv.org/abs/1511.06434
[52] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5046–5063, Sep. 2018.
[53] K. Sheng, W. Dong, Y. Kong, X. Mei, J. Li, C. Wang, F. Huang, and B.-G. Hu, "Evaluating the quality of face alignment without ground truth," *Comput. Graph. Forum*, vol. 34, no. 7, pp. 213–223, Oct. 2015.
[54] Z. Yang and T. Fang, "On the accuracy of image normalization by Zernike moments," *Image Vis. Comput.*, vol. 28, no. 3, pp. 403–413, Mar. 2010.
[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[56] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," Jan. 2017, *arXiv:1701.07875*. [Online]. Available: https://arxiv.org/abs/1701.07875
[57] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Nara, Japan, Apr. 1998, pp. 200–205.
[58] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. 3rd Int. Workshop CVPR Hum. Commun. Behav. Anal.*, San Francisco, CA, USA, Jun. 2010, pp. 94–101.
[59] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," 2017, *arXiv:1708.03985*. [Online]. Available: https://arxiv.org/abs/1708.03985
[60] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2584–2593.
[61] X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 318–335.
[62] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[63] S. Xie and H. Hu, "Facial expression recognition with FRR-CNN," *Electron. Lett.*, vol. 53, no. 4, pp. 235–237, Feb. 2017.
[64] S. Ghosh, A. Dhall, and N. Sebe, "Automatic group affect analysis in images via visual attribute and feature networks," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 1967–1971.
[65] Z. Li, "A discriminative learning convolutional neural network for facial expression recognition," in *Proc. 2nd IEEE Int. Conf. Comput. Commun. (ICCC)*, Chengdu, China, Dec. 2017, pp. 1641–1646.
[66] J. Deng, G. Pang, Z. Zhang, Z. Pang, H. Yang, and G. Yang, "cGAN based facial expression recognition for human-robot interaction," *IEEE Access*, vol. 7, pp. 9848–9859, Jan. 2019.
[67] X. Zhu, X. Zhang, X.-Y. Zhang, Z. Xue, and L. Wang, "A novel framework for semantic segmentation with generative adversarial network," *J. Vis. Commun. Image Represent.*, vol. 58, pp. 532–543, Jan. 2019.
[68] X. Zhu, Z. Li, X. Zhang, H. Li, Z. Xue, and L. Wang, "Generative adversarial image super-resolution through deep dense skip connections," *Comput. Graph. Forum*, vol. 37, no. 7, pp. 289–300, Oct. 2018.
[69] X.-Y. Zhang, H. Shi, X. Zhu, and P. Li, "Active semi-supervised learning based on self-expressive correlation with generative adversarial networks," *Neurocomputing*, vol. 345, pp. 103–113, Jun. 2019.
[70] L. Zhang, B. Verma, D. Tjondronegoro, and V. Chandran, "Facial expression analysis under partial occlusion: A survey," *ACM Comput. Surv.*, vol. 51, no. 2, Jun. 2018, Art. no. 25.
[71] L. Xu, H. Zhang, J. Raitoharju, and M. Gabbouj, "Unsupervised facial image de-occlusion with optimized deep generative models," in *Proc. 8th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Xi'an, China, Nov. 2018, pp. 1–6.
[72] R. A. Yeh, C. Chen, Y. L. Teck, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6882–6890.
[73] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "CVAE-GAN: Fine-grained image generation through asymmetric training," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2764–2773.

YANG LU received the B.S. degree from the College of Communication Engineering, Jilin University, Jilin, China, in 2014, where she is currently pursuing the Ph.D. degree. Her major research interests include pattern recognition and image processing.

SHIGANG WANG received the B.S. degree from Northeastern University, in 1983, the M.S. degree in communication and electronics from the Jilin University of Technology, in 1998, and the Ph.D. degree in communication and information system from Jilin University, China, in 2001, where he is currently a Professor. In recent years, he has authored or coauthored many papers and holds patents. His research interests include multidimensional signal processing and stereoscopic, and multi-view video coding. He was a recipient of many awards, including the China Institute of Communications Science and Technology Award, the Jilin Province Science and Technology Progress Award, and the Technology Invention Award. He is also the Director of the China Society of Image and Graphics and the Vice Chair of Jilin Province Society of Image and Graphics.

WENTING ZHAO received the B.S. degree from the College of Communication Engineering, Jilin University, Jilin, China, in 2012, where she is currently pursuing the Ph.D. degree. Her major research interest includes image processing.



YAN ZHAO received the B.S. degree in communication engineering from the Changchun Institute of Posts and Telecommunications, in 1993, the M.S. degree in communication and electronics from the Jilin University of Technology, in 1999, and the Ph.D. degree in communication and information system from Jilin University, in 2003, where she is currently a Professor. She was a Researcher with the Digital Media Institute, Tampere University of Technology, Finland, in 2003. In 2008, she was a Visiting Professor with the Vienna University of Technology. From 2013 to 2014, she was a Visiting Professor with the University of Ottawa, Canada. Her research interests include image and video processing, multimedia signal processing, and error concealment for audio and video transmitted over unreliable networks.